STEER-MOE: EFFICIENT AUDIO-LANGUAGE ALIGNMENT WITH A MIXTURE-OF-EXPERTS STEERING MODULE

Ruitao Feng[1] *Bixi Zhang*[2] *Sheng Liang*[1] *Zheng Yuan*[⋆]

[1] Independent Researcher, China
[2] The University of Hong Kong, Fauclty of Science, Hong Kong
[⋆] Aix-Marseille University, Laboratoire Parole et Langage (LPL), France

ABSTRACT

Aligning pretrained audio encoders and Large Language Models (LLMs) offers a promising, parameter-efficient path to building powerful multimodal agents. However, existing methods often require costly full-model finetuning or rely on static adapters that may lack expressive power. Drawing inspiration from the Platonic Representation Hypothesis, we introduce SteerMoE, a novel and modular framework for audio-language alignment. SteerMoE freezes both the audio encoder and the LLM decoder, training only a lightweight steering module integrated within the encoder's layers. This module uses a Mixture-of-Experts (MoE) router to dynamically select and apply learned steering vectors, progressively transforming continuous audio representations into a space comprehensible to the LLM. By operating entirely in the continuous embedding space, our approach requires no modifications to the LLM's vocabulary and preserves its advanced reasoning and agentic capabilities. We demonstrate through experiments on ASR, audio understanding, and a qualitative function-calling task that SteerMoE achieves strong performance while remaining highly modular and computationally efficient, offering a robust new paradigm for developing sophisticated audio-language systems.

***Index Terms*—** Multimodal Large Language Models, Parameter-Efficient Fine-Tuning (PEFT), Steering Vectors, Mixture-of-Experts (MoE), Platonic Hypothesis

1. INTRODUCTION

The human brain seamlessly integrates rich sensory inputs—sight, sound, and language—into a coherent model of the world. A central goal in artificial intelligence is to imbue machines with a similar capability, enabling them to reason across diverse data modalities. Recent theoretical work has formalized this pursuit under the Platonic Representation Hypothesis [1], which posits that neural networks, when trained on varied data, converge towards a shared, underlying representation of reality, much like Plato's Forms. According to this hypothesis, the projections of this universal reality onto different modalities (e.g., the visual appearance of a cat versus the sound of its meow) can be reconciled through simple, interpretable transformations [2, 3, 4].

This perspective offers a powerful paradigm for multimodal AI, suggesting that bridging the gap between modalities may not require monolithic, end-to-end training, but rather the discovery of efficient alignment functions within this shared latent space. In this work, we investigate this hypothesis within the audio-language domain, a critical nexus for human-computer interaction. We explore the challenge of aligning the continuous world of acoustics with the symbolic world of large language models (LLMs). Our contribution is a novel, parameter-efficient architecture that achieves this alignment through dynamic, context-aware steering, demonstrating a modular and effective path towards multimodal AI agents.

Current efforts to build audio-language models predominantly fall into two categories. The first involves training large, monolithic models from scratch or through extensive end-to-end finetuning on massive audio-text corpora [5, 6, 7, 8]. While these models achieve state-of-the-art performance, their development demands prohibitive computational resources, results in highly-coupled, inflexible architectures, and risks degrading the LLM's original reasoning capabilities. The second, more efficient paradigm involves parameter-efficient finetuning (PEFT), where a pretrained audio encoder is adapted to a frozen LLM. A common PEFT strategy is to discretize audio into a sequence of acoustic tokens, expanding the LLM's vocabulary to treat audio as another "language" [9, 10]. This, however, introduces architectural complexity via a separate quantizer, risks information loss, and compromises the LLM's modularity. A more direct approach uses a static adapter, such as a simple MLP, to map continuous audio features into the LLM's embedding space as a soft prompt [11, 12]. While parameter-efficient, such static mappings may lack the expressive power to perform the nuanced, context-dependent alignment required for diverse and complex speech tasks.

In this paper, we introduce SteerMoE, a framework that navigates a middle path between costly end-to-end training

and overly simplistic adaptation. Our approach directly operationalizes the Platonic Representation Hypothesis by learning a dynamic, content-aware alignment function. We insert a lightweight Mixture-of-Experts (MoE) module that operates internally within a frozen audio encoder. This module learns to select and apply a combination of expert "steering vectors" to the audio representations at each layer, adaptively modifying them to be seamlessly understood by a frozen LLM decoder. By manipulating representations directly in the continuous vector space, SteerMoE entirely bypasses audio tokenization, preserving the full richness of the acoustic signal and leaving the LLM's architecture untouched. This creates a truly 'plug-and-play' framework where components can be interchanged with ease. Our main contributions are:

- We introduce a dynamic, layer-wise steering mechanism based on MoE, offering a more expressive and parameter-efficient alignment than static adapters.
- We present a fully modular framework where audio encoders (e.g., Whisper, Conformer) and LLM decoders (e.g., Qwen, LLaMA) can be independently swapped, preserving their native reasoning and agentic capabilities.
- We provide strong experimental evidence that such a lightweight steering approach is sufficient to align audio and language representations by training and evaluating our model on diverse tasks, including Automatic Speech Recognition (ASR) on LibriSpeech and AISHELL-2, and audio understanding on Clotho-AQA benchmark.

2. METHODOLOGY

Our proposed framework, SteerMoE, is designed to be a modular and parameter-efficient solution for aligning audio and language representations. It consists of three primary components: a frozen pretrained audio encoder, a frozen pretrained LLM decoder, and a lightweight, trainable steering module that operates within the encoder's layers. The overall architecture is depicted in Figure 1.

2.1. Efficient Layer-wise Steering Module

The core of our method is a dynamic, layer-wise steering module that progressively refines the audio representations within the audio encoder. For an encoder with L layers, the module applies a content-aware adjustment at each layer. This module comprises a set of expert steering vectors, a shared router, a linear projection layer, and learnable scaling factors.

For each encoder layer $l \in \{1, ..., L\}$, we define a set of N learnable expert steering vectors, $\{E_{l,n}\}_{n=1}^{N}$, where each $E_{l,n} \in \mathbb{R}^D$ and D is the feature dimension of the encoder.

To maintain parameter efficiency, a single, shared MoE router, implemented as a linear layer with weights $W_{router} \in \mathbb{R}^{D \times (L \cdot N)}$, generates gating logits for all experts across all layers. Given the hidden state sequence $H_l \in \mathbb{R}^{T \times D}$ from the output of layer l, where T is the sequence length, the router computes gating scores $g_l \in \mathbb{R}^{T \times N}$ for that layer's experts as follows:

$$g_l = \text{softmax}(\text{slice}_l(H_l W_{router})) \quad (1)$$

where $\text{slice}_l(\cdot)$ is an operation that extracts the N logits corresponding to layer l. The resulting steering adjustment ΔH_l is the weighted sum of the expert vectors for that layer:

$$\Delta H_l = g_l E_l \quad (2)$$

This adjustment is then scaled by a learnable, per-layer parameter α_l and added back to the original hidden state to produce the final steered output H'_l:

$$H'_l = H_l + \alpha_l \Delta H_l \quad (3)$$

After the final steering operation at the L-th layer, we apply an average pooling layer with a kernel size of 4 across the temporal dimension. This downsampling step reduces the sequence length of the audio features, enhancing computational efficiency before they are passed to the LLM decoder. The resulting steered audio representation, H'_{audio}, is then interfaced with the LLM.

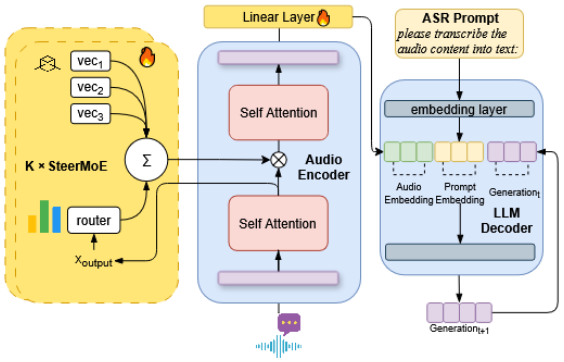

Fig. 1. The SteerMoE architecture. A frozen audio encoder processes the input waveform. At each layer, a trainable MoE steering module refines the audio representations. The final steered features are projected and used as a continuous prompt for a frozen LLM decoder.

2.2. Modality Alignment via Continuous Prompting

Our framework interfaces the steered audio representations with the LLM decoder by treating them as a continuous, or "soft", prompt. This approach operates entirely in the continuous vector space, avoiding the information loss and architectural modifications associated with discrete audio tokenization.

The final sequence of steered audio vectors from the encoder, $H'_{audio} \in \mathbb{R}^{T_{audio} \times D}$, is first passed through a trainable linear projection layer with weights $W_{proj} \in \mathbb{R}^{D \times D_{llm}}$

to match the LLM's hidden dimension, D_{llm}. This results in a sequence of audio prompt embeddings P_{audio}. These embeddings are then prepended to the standard text embeddings E_{text} to form the final input sequence for the LLM decoder:

$$E_{input} = [H'_{audio}W_{proj}; E_{text}] = [P_{audio}; E_{text}] \quad (4)$$

2.3. Training Objective

The trainable parameters of SteerMoE—comprising the expert steering vectors $\{E_{l,n}\}$, the shared router W_{router}, the scaling factors $\{\alpha_l\}$, and the projection layer W_{proj}—are optimized using a standard auto-regressive, next-token prediction objective. The model learns to predict the next token in the text sequence, conditioned on the audio prompt and the preceding text tokens.

Crucially, the cross-entropy loss is computed *only* over the target text tokens. The audio prompt portion of the input sequence is masked out from the loss calculation. Given a target text sequence $Y = \{y_1, ..., y_{T_{text}}\}$, the objective is to minimize the negative log-likelihood:

$$\mathcal{L}(\theta) = -\sum_{t=1}^{T_{text}} \log p(y_t | P_{audio}, y_{<t}; \theta) \quad (5)$$

where θ represents all trainable parameters. This ensures that the steering module learns to transform audio representations into a format that the frozen LLM can effectively use to condition its text generation.

3. EXPERIMENTS

We evaluate our SteerMoE framework on two distinct task categories: foundational **Automatic Speech Recognition (ASR)** and complex **Audio Question Understanding (AQU)**.

3.1. Datasets and Tasks

To measure the core quality of the audio-to-text alignment, we use two standard ASR benchmarks:

- **LibriSpeech** [13]: A public domain corpus for English speech recognition, derived from audiobooks.
- **AISHELL-2** [14]: A 1000-hour corpus for Mandarin Chinese ASR, recorded in a quiet indoor environment.

To evaluate the model's ability to perform complex reasoning over audio, we use the Clotho-AQA benchmark.

- **Clotho-AQA** [15] is a publicly available audio question answering dataset consisting of 1,991 audio samples of 15–30 seconds each, drawn from the Clotho dataset. Each sample has six questions, and for each question, three different annotators provide answers, yielding a total of 35,838 question-answer pairs. Four questions per sample are binary (yes/no) and two are single-word answer questions.

3.2. Implementation Details

Our SteerMoE model is built upon powerful, publicly available pretrained components. We use the Whisper-large-v3 model as our frozen audio encoder and the Qwen2.5-7B-Instruct model as our frozen LLM decoder. The only trainable parameters are those within our steering module and the final projection layer. The steering module employs $N = 8$ experts with an initial steering scale $\alpha_l = 0.1$. Training is performed with a batch size of 4, AdamW optimizer, FP16 precision, and separate learning rates for the base model (1×10^{-4}), steering vectors (1×10^{-2}), and router (1×10^{-3}).

Audio inputs are standardized to 16kHz mono, and log-Mel spectrogram features are extracted using the Whisper feature extractor. Text transcriptions are tokenized with the Qwen tokenizer, prepended with an instructional prompt: "*please transcribe the audio content into text:* " for English ASR task, "请逐字复述音频内容为文字" (Please transcribe the audio input word by word) for Chinese ASR task, and "*lease answer the following question. The question is :* " + batch["Question Text"] for audio understanding task. We filter samples longer than 30 seconds or 448 text tokens due to the limitation of computational resources. A custom data collator handles padding and masks the audio prompt tokens from the loss calculation.

3.3. Main Results

We compare SteerMoE against strong published results for representative baseline models on both ASR and AQU tasks.

3.3.1. ASR Performance

As shown in Table 1, our parameter-efficient SteerMoE model achieves competitive ASR performance. To explicitly validate the "plug-and-play" modularity of our approach, we tested two different frozen audio encoders: Whisper-large-v3 and Conformer. While not yet matching the performance of a fully-finetuned system, both configurations demonstrate strong transcription capabilities by training only a minuscule fraction of the total parameters, validating the effectiveness of our alignment strategy.

Table 1. ASR results (CER/WER %) on LibriSpeech and AISHELL-2 test sets. Lower is better. W and C in the model name indicate the Whisper-large-v3 or Conformer encoder applied and the number after its parameter size

Model	LS (WER)	AS-2 (CER)
Whisper-large-v3	2.7	4.96
SteerMoE (W7B)	5.69	5.96
SteerMoE (C3B)	3.26	3.44
SteerMoE (C7B)	**2.42**	**2.50**

3.3.2. Audio Understanding Performance

On the challenging ClothoAQA benchmark, SteerMoE demonstrates its ability to unlock the LLM's powerful reasoning capabilities for spoken inputs. As shown in Table 2, our model performs favorably against strong, larger multimodal models, highlighting the benefits of preserving the LLM's integrity through our frozen-decoder approach.

Table 2. Performance on the ClothoAQA benchmark (Average Accuracy %). Higher is better. Includes both industrial multimodal LLMs and our SteerMoE models. Data for the industrial models taken from `https://github.com/MoonshotAI/Kimi-Audio`.

Model	Param.	Avg. Acc. (%)
Kimi-Audio	9.77B	71.24
Step-Audio-Chat	130B	45.84
SteerMoE (W7B)	**7B + 1.5B + 64M**	**52.35**
SteerMoE (C3B)	3B + 1.5B + 64M	46.24
SteerMoE (C7B)	7B + 1.5B + 64M	49.06

3.3.3. Qualitative Analysis of Agentic Capabilities

A key claim of our work is that by freezing the LLM, its native capabilities are preserved. To test this, we conducted a simple qualitative experiment to probe the model's agentic function-calling ability. We configured a tool-use scenario where a specific spoken query, "上海的天气怎么样?" (What's the weather in Shanghai?), should trigger a predefined function that returns a non-literal, absurd answer. When presented with the audio of this query, our SteerMoE model correctly interpreted the user's intent from speech and successfully triggered the function call, yielding the predefined response: "上海今天下黄金" (It's raining gold in Shanghai today). This successful outcome demonstrates that the complex machinery for tool-use and function calling, inherent to the frozen Qwen LLM, remains fully intact and is directly accessible via spoken commands through our alignment module.

3.4. Ablation Studies

To validate our architectural design, we conducted several ablation studies, with results summarized in Table 3. First, we replaced our dynamic MoE-based steering module with a simple linear projection layer (a static adapter). The significant performance drop underscores the importance of a dynamic, content-aware alignment mechanism. Second, we experimented with varying the number of expert vectors per layer, finding that using 8 experts provides a strong balance of performance and parameter efficiency compared to 2 or 4 experts. These results confirm that the expressiveness of the MoE router is a critical component of our model's success.

Table 3. Ablation studies on LibriSpeech.

Model Variant	WER (%)
SteerMoE (8 Experts)	**2.42%**
SteerMoE (4 Experts)	3.10%
SteerMoE (2 Experts)	6.22%
Static Adapter (No MoE)	103%

4. DISCUSSION

Our results demonstrate that a lightweight, dynamic steering module can effectively align the representational spaces of separate, pretrained audio and language models. The success of this parameter-efficient approach, which operates entirely in the continuous embedding space, provides strong empirical support for the Platonic Representation Hypothesis in the audio-language domain. It suggests that complex, deep fusion is not a prerequisite for multimodal understanding; rather, discovering the correct transformations within a shared conceptual space is sufficient.

The primary advantage of our method is its modularity and preservation of the LLM's inherent capabilities. By freezing the decoder and avoiding vocabulary modification, SteerMoE makes the core components "plug-and-play" and, more importantly, unlocks the LLM's advanced reasoning and agentic abilities for spoken input, as shown in our function-calling experiments. This positions our work not just as an ASR system, but as a general framework for building sophisticated audio-language agents.

Nonetheless, we acknowledge several limitations. Our experiments were conducted on relatively clean speech corpora. The model's robustness in highly noisy environments and its performance on a wider diversity of non-speech audio remain open questions. Furthermore, while our training was constrained by data scale and sequence length due to available computational resources, our architectural design theoretically allows the model at inference to process audio sequences of any length that the underlying LLM decoder's context window can accommodate. Our agentic experiment, while a successful proof-of-concept, also requires testing on more complex, multi-turn interactions.

For future work, a particularly exciting direction is to probe the learned expert steering vectors. Analyzing whether specific experts come to specialize in phonetic features, prosody, speaker identity, or even noise separation could yield fascinating insights into the nature of the learned audio-language mapping.

5. COMPLIANCE WITH ETHICAL STANDARDS

This research study was conducted retrospectively using only human-subject data made available in open access by the respective dataset providers stated in the original papers cited (LibriSpeech, AISHELL-2, and Clotho-AQA). Ethical approval was not required as confirmed by the licenses attached with the open access data.

6. REFERENCES

[1] Minyoung Huh, Brian Cheung, Tongzhou Wang, and Phillip Isola, "Position: the platonic representation hypothesis," in *Proceedings of the 41st International Conference on Machine Learning*. 2024, ICML'24, JMLR.org.

[2] Youcheng Huang, Chen Huang, Duanyu Feng, Wenqiang Lei, and Jiancheng Lv, "Cross-model transferability among large language models on the platonic representations of concepts," in *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics*, Vienna, Austria, July 2025, pp. 3686–3704.

[3] Yusong Wu, Ke Chen, Tianyu Zhang, Yuchen Hui, Taylor Berg-Kirkpatrick, and Shlomo Dubnov, "Large-scale contrastive language-audio pretraining with feature fusion and keyword-to-caption augmentation," in *2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, June 2023, pp. 1–5.

[4] Megan Tjandrasuwita, Chanakya Ekbote, Liu Ziyin, and Paul Pu Liang, "Understanding the emergence of multimodal representation alignment," in *Forty-second International Conference on Machine Learning*, 2025.

[5] Yunfei Chu, Jin Xu, Qian Yang, Haojie Wei, Xipin Wei, Zhifang Guo, Yichong Leng, Yuanjun Lv, Jinzheng He, Junyang Lin, et al., "Qwen2-audio technical report," *arXiv preprint arXiv:2407.10759*, 2024.

[6] KimiTeam, Ding Ding, Zeqian Ju, Yichong Leng, Songxiang Liu, Tong Liu, Zeyu Shang, Kai Shen, Wei Song, Xu Tan, Heyi Tang, Zhengtao Wang, Chu Wei, Yifei Xin, Xinran Xu, Jianwei Yu, Yutao Zhang, Xinyu Zhou, Y. Charles, Jun Chen, Yanru Chen, Yulun Du, Weiran He, Zhenxing Hu, Guokun Lai, Qingcheng Li, Yangyang Liu, Weidong Sun, Jianzhou Wang, Yuzhi Wang, Yuefeng Wu, Yuxin Wu, Dongchao Yang, Hao Yang, Ying Yang, Zhilin Yang, Aoxiong Yin, Ruibin Yuan, Yutong Zhang, and Zaida Zhou, "Kimi-audio technical report," 2025.

[7] Kai-Tuo Xu, Feng-Long Xie, Xu Tang, and Yao Hu, "Fireredasr: Open-source industrial-grade mandarin speech recognition models from encoder-decoder to llm integration," *arXiv preprint arXiv:2501.14350*, 2025.

[8] Weiguo Wang, Andy Nie, Wenrui Zhou, Yi Kai, and Chengchen Hu, "Teaching physical awareness to LLMs through sounds," in *Forty-second International Conference on Machine Learning*, 2025.

[9] Zuhair Hasan Shaik, Pradyoth Hegde, Prashant Bannulmath, and Deepak K T, "LaRA: Large rank adaptation for speech and text cross-modal learning in large language models," in *Findings of the Association for Computational Linguistics: EMNLP 2024*, Miami, Florida, USA, Nov. 2024, pp. 8201–8211.

[10] Heeseung Kim, Soonshin Seo, Kyeongseok Jeong, Ohsung Kwon, Soyoon Kim, Jungwhan Kim, Jaehong Lee, Eunwoo Song, Myungwoo Oh, Jung-Woo Ha, Sungroh Yoon, and Kang Min Yoo, "Paralinguistics-aware speech-empowered large language models for natural conversation," in *Advances in Neural Information Processing Systems*, 2024, vol. 37, pp. 131072–131103.

[11] Fawaz Sammani and Nikos Deligiannis, "Zero-shot natural language explanations," in *International Conference on Representation Learning*, Y. Yue, A. Garg, N. Peng, F. Sha, and R. Yu, Eds., 2025, vol. 2025, pp. 85084–85107.

[12] Shujie Hu, Long Zhou, Shujie Liu, Sanyuan Chen, Lingwei Meng, Hongkun Hao, Jing Pan, Xunying Liu, Jinyu Li, Sunit Sivasankaran, Linquan Liu, and Furu Wei, "WavLLM: Towards robust and adaptive speech large language model," in *Findings of the Association for Computational Linguistics: EMNLP 2024*, Miami, Florida, USA, Nov. 2024, pp. 4552–4572.

[13] Vassil Panayotov, Guoguo Chen, Daniel Povey, and Sanjeev Khudanpur, "Librispeech: an asr corpus based on public domain audio books," in *2015 IEEE international conference on acoustics, speech and signal processing (ICASSP)*. IEEE, 2015, pp. 5206–5210.

[14] Jiayu Du, Xingyu Na, Xuechen Liu, and Hui Bu, "Aishell-2: Transforming mandarin asr research into industrial scale," *arXiv preprint arXiv:1808.10583*, 2018.

[15] Samuel Lipping, Parthasaarathy Sudarsanam, Konstantinos Drossos, and Tuomas Virtanen, "Clotho-aqa: A crowdsourced dataset for audio question answering," in *2022 30th European Signal Processing Conference (EUSIPCO)*, 2022, pp. 1140–1144.